DA 204o: Data Science in Practice
Course Project Proposal

Forecasting Urban Air Quality for Public Health Advisories

- Shivam Kumar, kshivam@iisc.ac.in
- Sachin R, sachinr@iisc.ac.in
- Prakash S, prakashs1@iisc.ac.in
- Pothukanuri Sai Venkat, saivenkatp@iisc.ac.in

Image source: Internet

Problem Statement and Motivation

- Background of the problem
  - Major Indian cities frequently experience severe air pollution events, posing significant public health risks. Citizens and authorities lack a reliable, forward-looking tool to anticipate these events.
- Why is it important?
  - Accurate forecasts allow public health agencies to issue timely warnings, especially for vulnerable populations (children, elderly, asthmatics). It also empowers individuals to take preventive measures and helps policymakers evaluate the short-term effectiveness of anti-pollution initiatives.
- Objectives of the project
  - To build and evaluate a robust time-series forecasting model capable of predicting the daily average Air Quality Index (AQI) for a major Indian city (e.g., Delhi) 24 to 72 hours in advance.

WHO PM2.5 (µg/m³) Recommended Guidelines as of 2024: **0-5.0**

AQI Level		PM2.5 (µg/m³)	Health Recommendation (for 24 hour exposure)
Good	0-50	0-9.0	Air quality is satisfactory and poses little or no risk.
Moderate	51-100	9.1-35.4	Sensitive individuals should avoid outdoor activity as they may experience respiratory symptoms.
Unhealthy for Sensitive Groups	101-150	35.5-55.4	General public and sensitive individuals in particular are at risk to experience irritation and respiratory problems.
Unhealthy	151-200	55.5-125.4	Increased likelihood of adverse effects and aggravation to the heart and lungs among general public.
Very Unhealthy	201-300	125.5-225.4	General public will be noticeably affected. Sensitive groups should restrict outdoor activities.
Hazardous	301+	225.5+	General public at high risk of experiencing strong irritations and adverse health effects. Should avoid outdoor activities.

A boy plays cricket on a ground near the banks of Yamuna river, on a smoggy morning, amid ongoing air pollution in New Delhi, India, on Thursday. Delhi and its surrounding areas are blanketed by thick smog every winter. Photo: Reuters

AQI AT 5PM

RK Puram	422	Narela	398
New Moti Bagh	421	DU North Campus	397
Wazirpur	421	Alipur	392
Bawana	420	Major Dhyan Chand National Stadium	392
Karni Singh Shooting Range	419	Burari Crossing	378
Sonia Vihar	414	Sri Aurobindo Marg	378
Pusa	411	Ashok Vihar	369
Okhla Phase-2	409	Lodhi Road	360
Shadipur	406	Vivek Vihar	327
Mandir Marg	402	CRRI Mathura Road	315
Najafgarh	402	DTU	306
Jawaharlal Nehru Stadium	401	IHBAS, Dilshad Garden	222

Data Overview

- Datasets – city_day (daily data), city_hour (hourly data)

city_day

City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Delhi	2015-01-01	313.22	607.98	69.16	36.39	110.59	33.85	15.2	9.25	41.68	14.36	24.86	9.84	472.0	Severe
Delhi	2015-01-02	186.18	269.55	62.09	32.87	88.14	31.83	9.54	6.65	29.97	10.55	20.09	4.29	454.0	Severe
Delhi	2015-01-03	87.18	131.9	25.73	30.31	47.95	69.55	10.61	2.65	19.71	3.91	10.23	1.99	143.0	Moderate
Delhi	2015-01-04	151.84	241.84	25.01	36.91	48.62	130.3	11.54	4.63	25.36	4.26	9.71	3.34	319.0	Very Poor
Delhi	2015-01-05	146.6	219.13	14.01	34.92	38.25	122.8	9.2	3.33	23.2	2.8	6.21	2.96	325.0	Very Poor
Delhi	2015-01-06	149.58	252.1	17.21	37.84	42.46	134.9	9.44	3.66	26.83	3.63	7.35	3.47	318.0	Very Poor
Delhi	2015-01-07	217.87	376.51	26.99	40.15	52.41	134.8	9.78	5.82	28.96	4.93	9.42	5.21	353.0	Very Poor
Delhi	2015-01-08	229.9	360.95	23.34	43.16	51.21	138.1	11.01	3.31	30.51	5.8	11.4	4.83	383.0	Very Poor

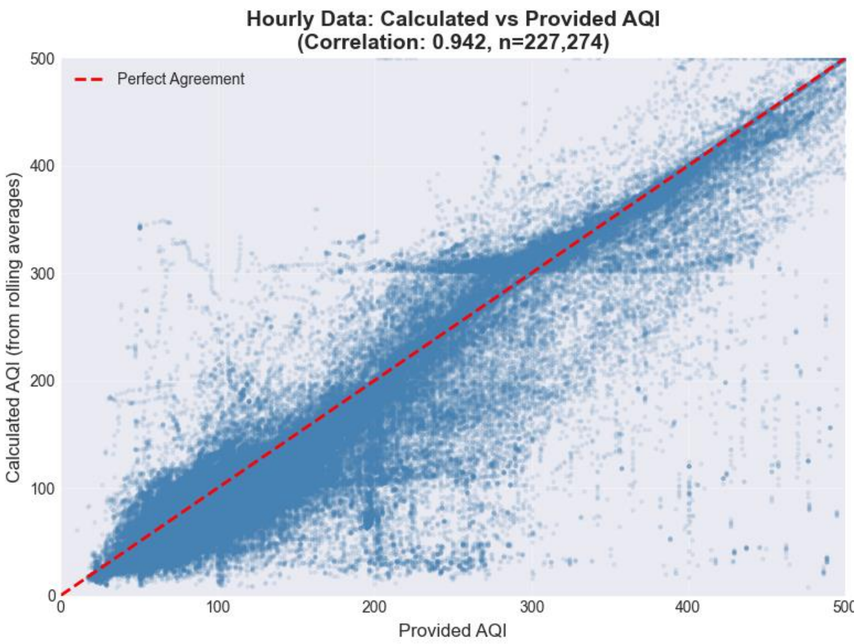
city_hour

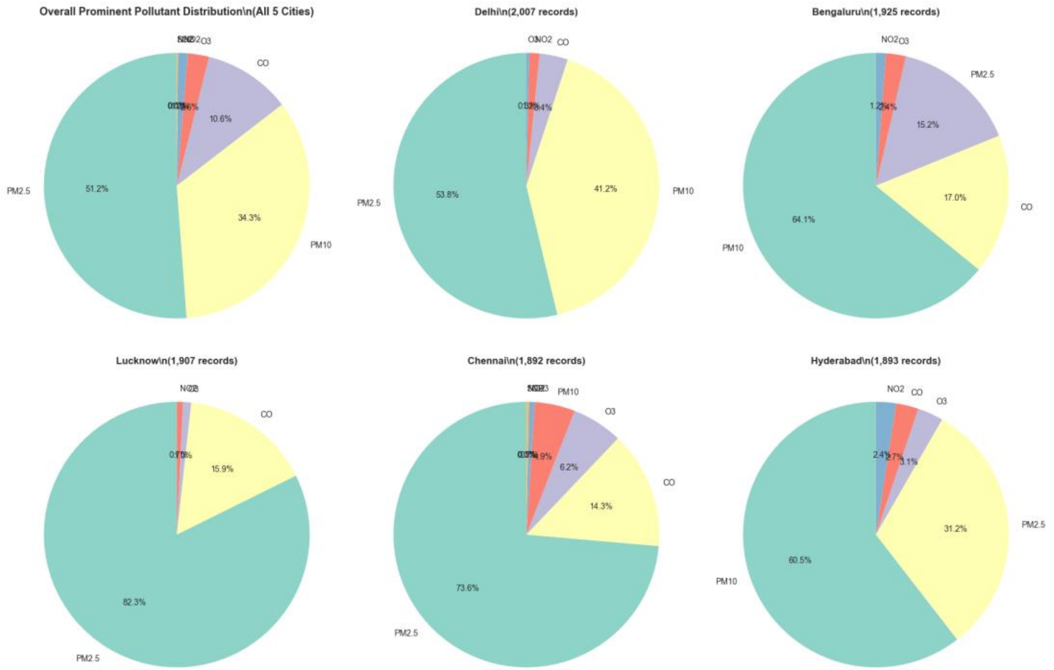
City	Datetime	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Delhi	2015-01-01 01:00:00	454.58	935.18	81.52	41.78	187.66	27.54	9.29	3.41	54.94	25.24	58.57	13.8		
Delhi	2015-01-01 02:00:00	440.44		70.8	43.46	176.83	27.72	13.28	3.88	50.53	23.1	49.37	15.63		
Delhi	2015-01-01 03:00:00	409.09		132.4	41.19	141.02	28.94	29.67	2.83	19.33	19.04	38.94	17.18		
Delhi	2015-01-01 04:00:00	436.12		84.78	39.55	102.84	29.3	21.76	4.33	20.08	13.99	27.53	16.82		
Delhi	2015-01-01 05:00:00	415.88	976.99	60.24	37.41	80.12	30.84	26.19	6.17	16.0	11.14	21.99	14.29		
Delhi	2015-01-01 06:00:00	384.16	862.23	59.84	32.06	78.34	30.71	11.04	7.33	12.33	10.7	20.85	12.42		
Delhi	2015-01-01 07:00:00	344.44	731.83	66.55	30.97	84.67	30.64	8.39	8.0	58.67	10.15	19.8	10.35		

EDA – CPCB AQI calculation

- CPCB: Central Pollution Control Board
- Subindex are calculated using for each pollutant based on CPCB guidelines
  - AQI = Max of subindices
- Gained insight into dominant pollutants (PM2.5, PM10, CO)

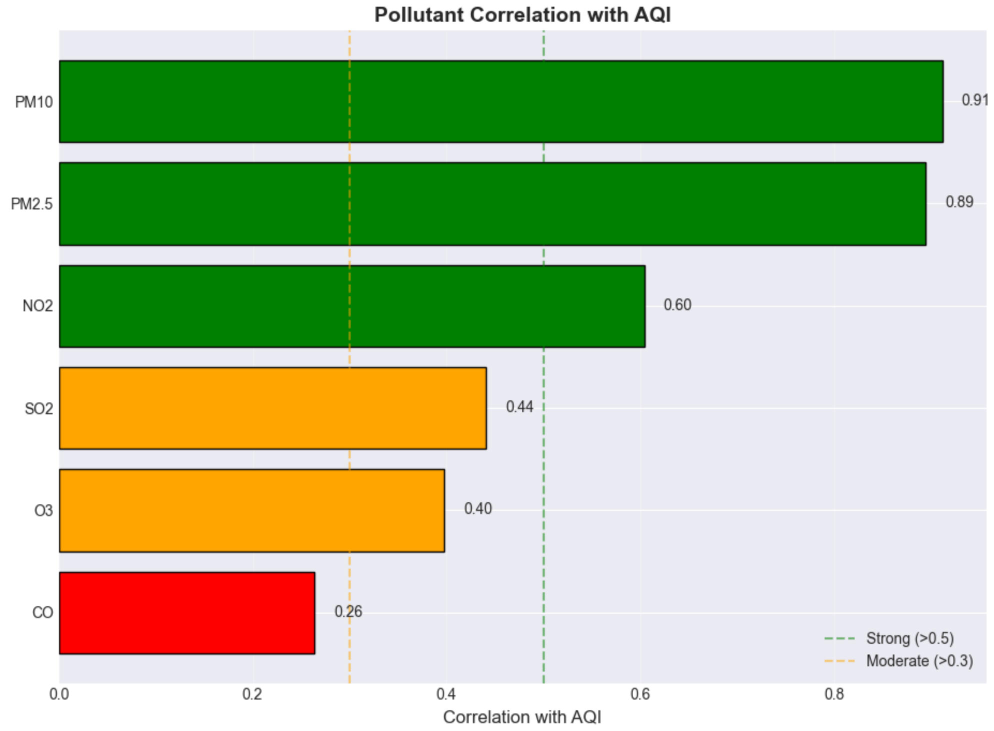
	PM10 (ug/m3)	PM2.5 (ug/m3)
0-50	0-50	0-30
51-100	51-100	31-60
101-200	101-250	61-90
201-300	251-350	91-120
301-400	351-430	121-250
401-500	430+	250+

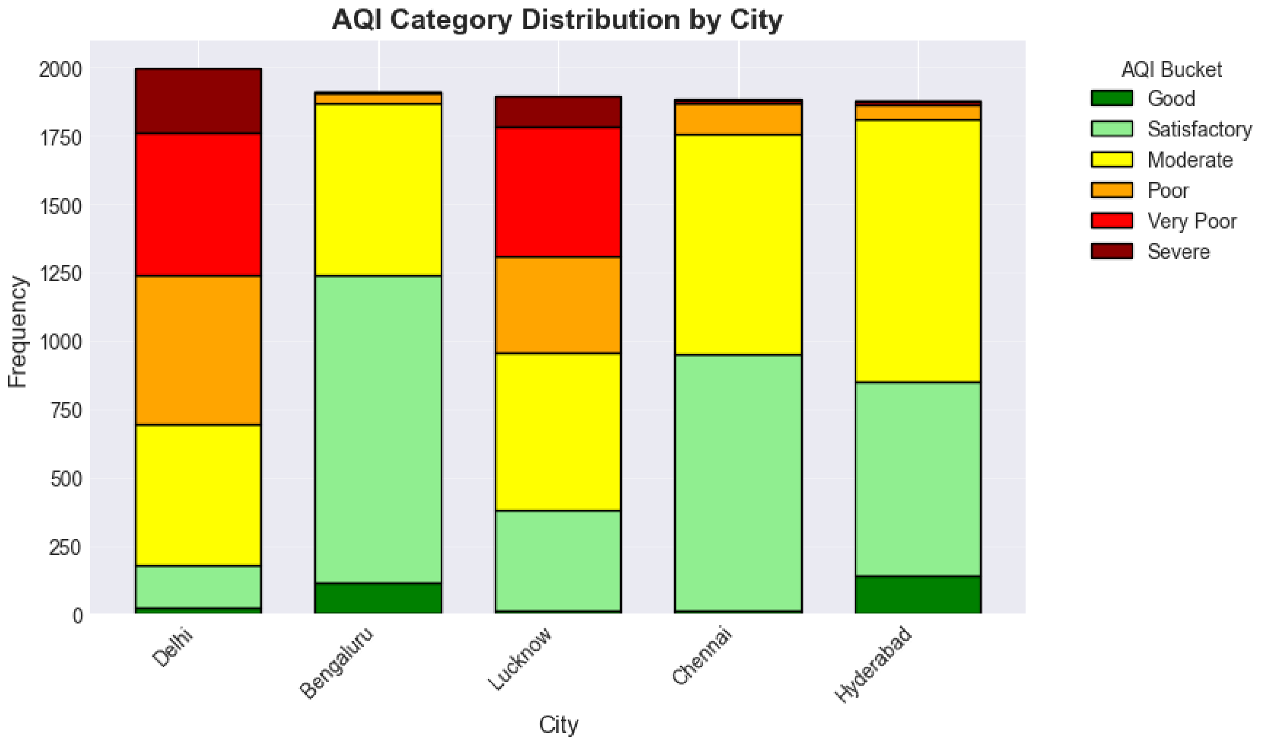




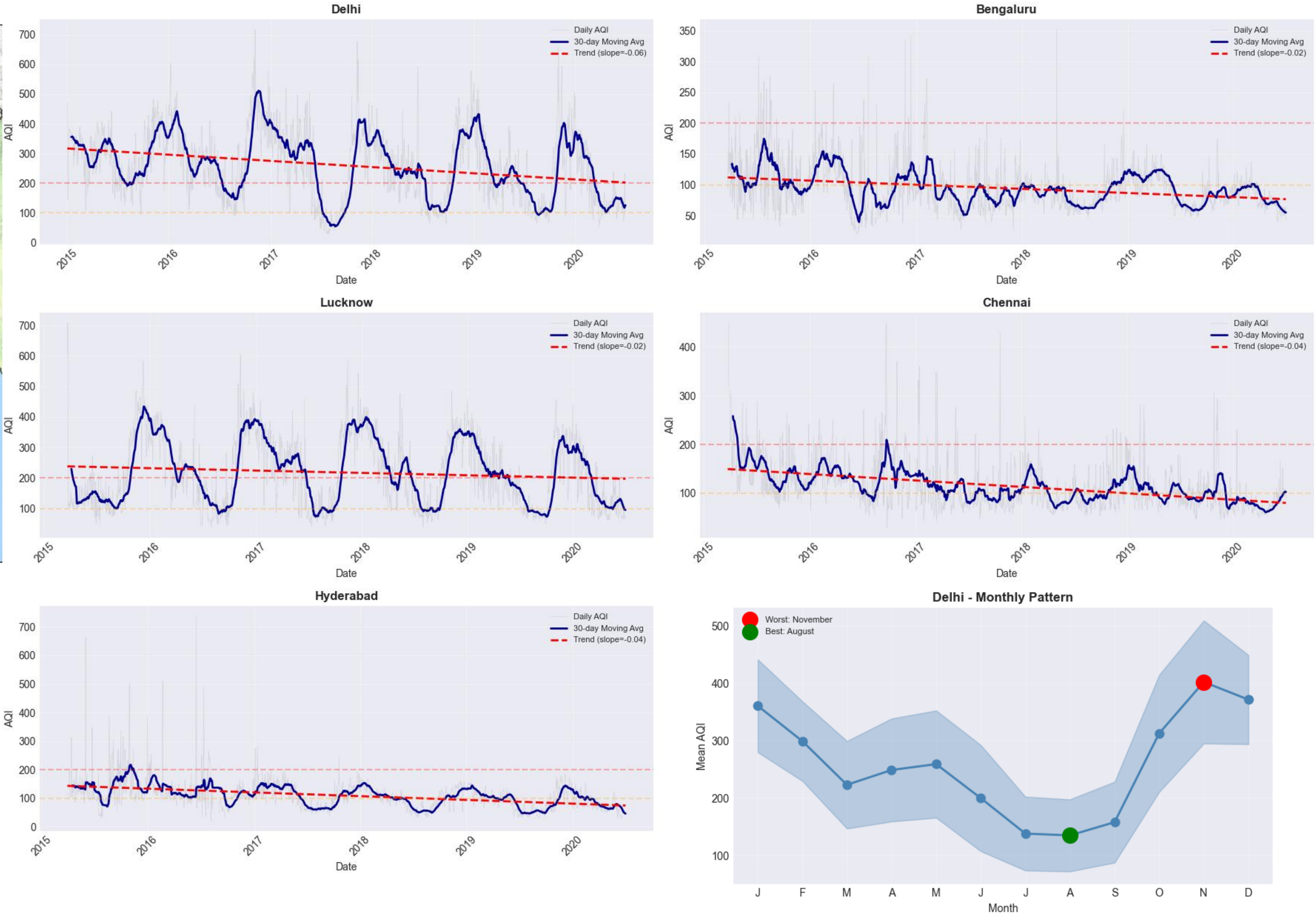
EDA: Univariate, Bivariate and Multivariate analysis

- All pollutants have right skewed distribution with varying degree of skewness
- Pollutant correlation with AQI matched with expected CPCB calculations
- Comparison among cities: Delhi had the worst AQI results





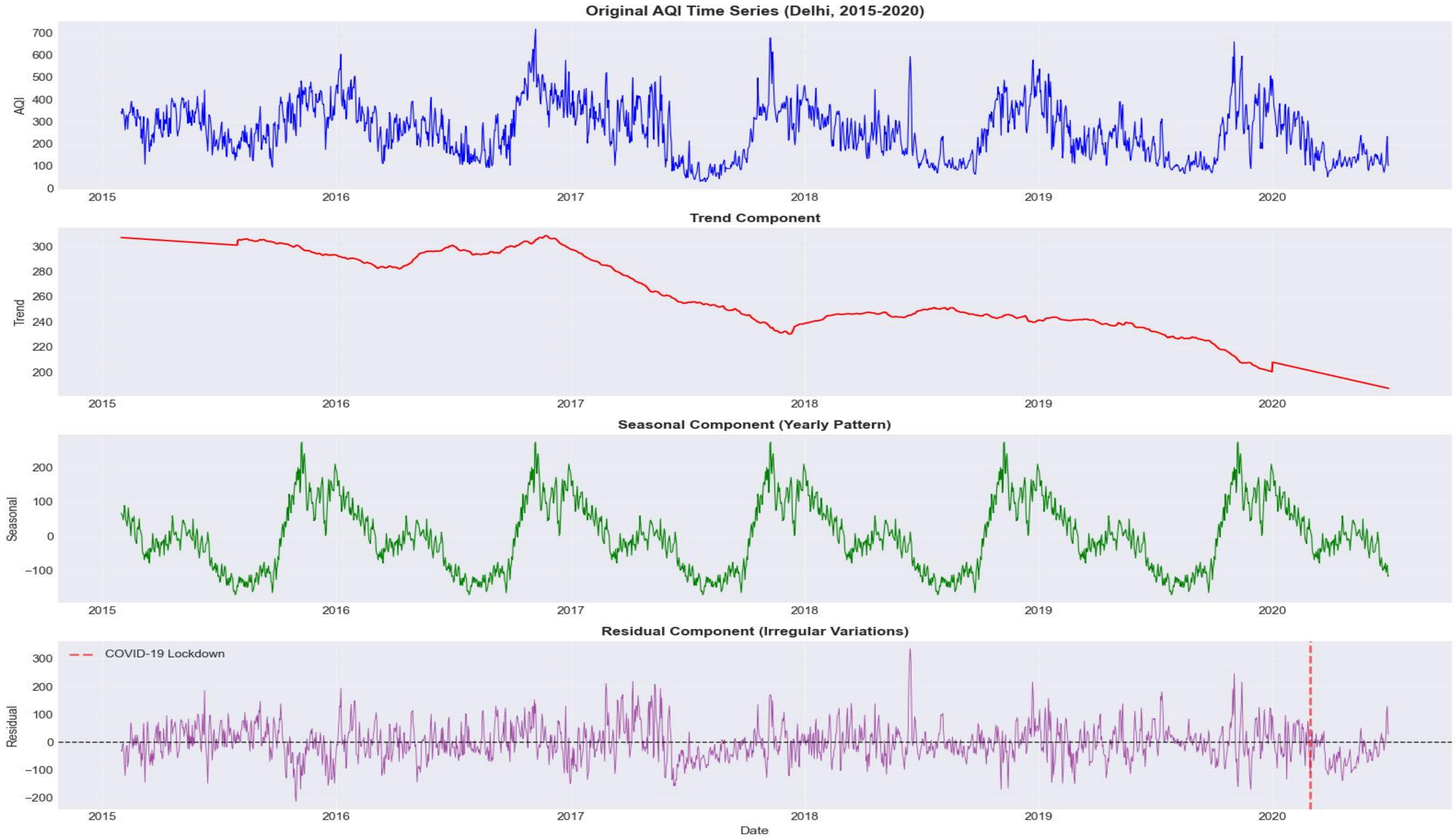
EDA – Temporal analysis



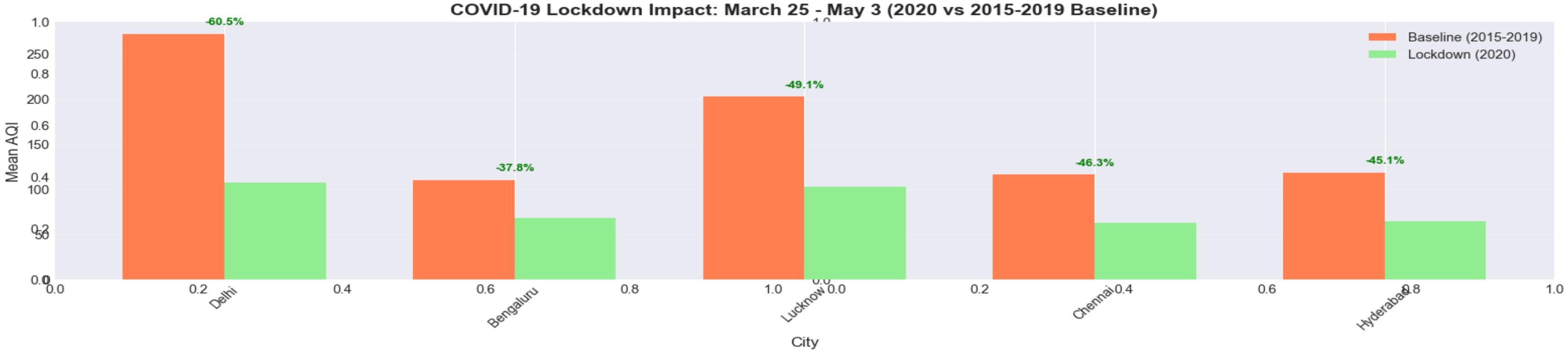
- Strong seasonal patterns but patterns differs based on city
- Similar patterns for cities with geographically closer to each other

EDA – Time series decomposition (Delhi)

Data passed the Augmented Dickey-Fuller (ADF) test for stationarity (p-value = 0.011 < 0.05)



EDA – Impact of COVID



Data Preparation

Missing value imputation

- Imputing missing pollutant data:
  - Forward fill for short gaps
  - Month median from other years for long gaps
  - Backward fill for edge cases
- Imputing missing AQI:
  - CPCB formula

Feature engineering

- Decomposed date feature into year, month, day and hour
- Added lag features for AQI and pollutants (t-1, t-2, t-3, t-7 …)
- Added rolling window features for AQI and pollutants
- Handled missing values

Train-Val-Test split

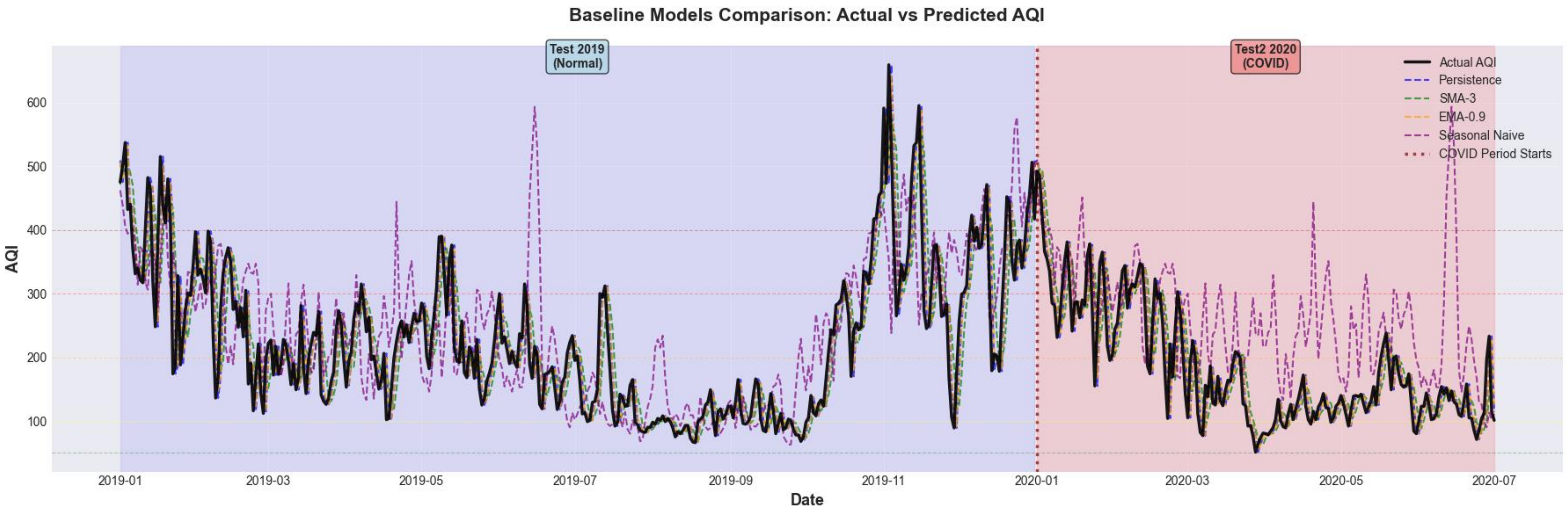
- Train-Val-Test split as below:
- Train: 2015-2017
- Val: 2018
- Test1: 2019
- Test2: 2020
- Created two different test set to capture impact due to COVID

Naïve Baseline models

- Persistence :
  - No parameter tuning
  - Performs well when Pollution levels usually do not change drastically (AQI shows high day-to-day autocorrelation)
- Simple Moving Average (SMA) :
  - window size Small k (SMA-3 / SMA-7)
  - Reacts faster to recent AQI changes
  - Better RMSE than large windows
- Exponential Moving Average (EMA) :
  - High α (EMA-0.9) → best performance among EMA variants
  - Closely tracks sudden AQI jumps
  - Works well in periods of high pollution volatility
- Seasonal Naïve (365-day Lag)
  - Works well when the pollution cycle repeats yearly
  - Fails during extraordinary periods (COVID lockdown)

Naïve Baseline models

Model	MAE	RMSE	MAPE	R2	Category_Accuracy
Persistence	38.48	53.99	17.72%	0.79	64.66%
EMA-0.9	38.87	54.38	18.01%	0.79	64.38%
SMA-3	48.49	66.06	23.09%	0.68	55.07%
Seasonal_Naive-365	70.17	95.05	37.05%	0.34	41.64%



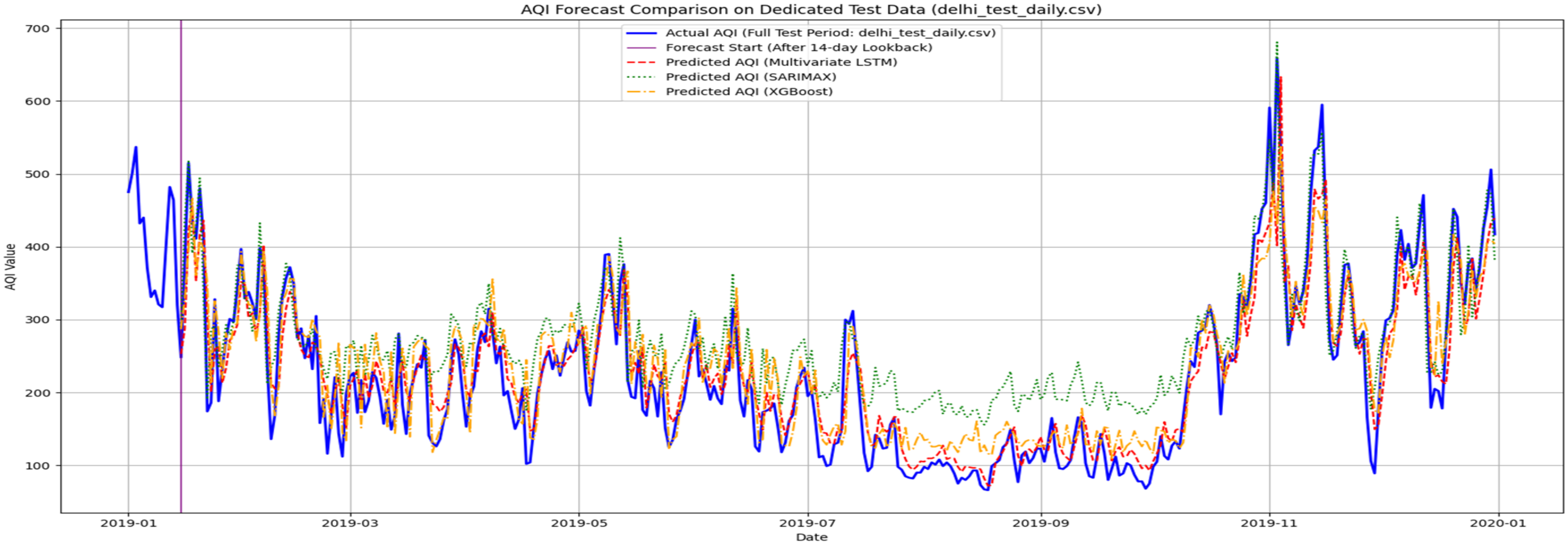
Advanced Models

SARIMA : Captures seasonal patterns, models trend + seasonality + autocorrelation.
LSTM (Long Short-Term Memory) : Deep-learning model for sequence data.
XGBoost (Extreme Gradient Boosting) : State-of-the-art tree-based model. Excellent for tabular time-series

Model	Parameter	Value
LSTM	LSTM Units	128 (per layer)
	Stacked Layers	2
	Dropout Rate	0.3
	Look-Back Window	14 days
XGBoost	Max Depth	7
	Learning Rate	0.03
	N Estimators	1000
	Feature Count	13
SARIMAX	(p, d, q)	(1, 1, 1)
	(P, D, Q, m)	(1, 1, 1, 7)
	Exogenous Count	12

Plots & Results - 2019

Model	MAE	MAPE	RMSE	R2	Category Accuracy
SARIMAX	54.84	38.73%	63.06	0.6899	47.01%
LSTM	32.90	17.14%	44.07	0.8485	68.38%
XGBoost	39.65	22.96%	51.78	0.7909	59.83%



Plots & Results - 2020

Model	MAE	MAPE	RMSE	R2	Category Accuracy
XGBoost	36.37	27.13%	43.89	0.6734	65.09%
LSTM	29.78	22.17%	38.12	0.7536	73.96%
SARIMAX	69.15	57.42%	76.21	0.0151	34.32%